HOW TO WRITE THESES WITH TWO LINE TITLES

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

John Henry Candidate
January 2024

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

14

(John Parker) Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

15

(John Green)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

16

(John BigBooty)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

17

(Jane Supernumerary)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

18

(Severus Snape)

Approved for the Stanford University Committee on Graduate Studies

19

Preface

This thesis tells you all you need to know about...

Acknowledgments

I would like to thank...

Contents

List of Tables

List of Figures

Chapter 1

Introduction

> The major challenge for language documentation in the next decade or two is what could be called the transcription challenge. This is a multilayered challenge that goes far beyond the practical challenge of speeding up the transcription process. [...] Despite its centrality to language documentation, transcription remains critically undertheorized and understudied. Further progress in language documentation, and ultimately also its overall success, crucially depends on further investigating and understanding the transcription process, broadly conceived.
>
> ---
>
> Himmelmann, N. P. (2018). Meeting the transcription challenge. In B. McDonnell, A. Berez-Kroeker & G. Holton (Eds.), *Reflections on language documentation: 20 years after Himmelmann 1998*. University of Hawai'i Press.

Language documentation is typically considered a sub-field of linguistics and its purpose is "to provide a comprehensive record of the linguistic practices characteristic of a given speech community" (Himmelmann, 1998, p. 166). Many aspects of its practice, however, require continual integration of developments beyond linguistics. As the target languages are often spoken by minoritised Indigenous communities, it is necessary to proactively consider evolving norms regarding fair and ethical collaboration (Holton et al., 2022). Concurrently, data protocols must not only continually integrate these norms but also technological advancements that more effectively enable the creation, processing, storage, and distribution of recorded materials (Berez-Kroeker et al., 2023). In both respective areas, there are ongoing developments that constitute paradigm shifts: 1) decolonising, community-centred approaches to working with Indigenous communities, and 2) speech and language processing systems powered by foundation models (FMs). Integrating considerations from the former and leveraging advancements in the latter, I examine in this dissertation how these FM-powered systems can help create contextually-appropriate solutions to combat a major bottleneck in language documentation workflows — the transcription of recorded materials.

1.1 Language documentation and the transcription bottleneck

It is generally recognised that there are about 7,000 languages spoken in the world today and that at least half of them may not exist by the end of the century (Austin & Sallabank, 2011). Many of these languages are spoken by Indigenous and minoritised communities, who are under cultural, economic, and technological pressures to shift to using more dominant regional, national, or global languages. For example, education may be conducted entirely in more widely-used regional or national languages and technological interfaces may only exist in a national or global language. Such pressures can lead to language endangerment, as younger generations gradually adopt the more widely-used languages until there remain no living first-language or 'L1' speakers, at which point a language may become 'dormant'.

There are many ongoing revitalisation efforts to fight these pressures by encouraging the learning and active use of endangered languages, safeguard the knowledge of elder L1 speakers by recording them, as well as efforts to 'awaken' dormant ones based on archival recordings. However, as recording speech is considerably easier than transcribing the recorded speech, many recording collections of these languages remain only partially transcribed, if at all (Cox et al., 2019). Yet, raw audio data comprising untranscribed speech is difficult to index and search, limiting how efficiently content within the materials can be discovered and used (e.g. for creating language learning materials). Indeed, the difficulty is so ubiquitous that it has been termed the "transcription bottleneck" (Foley et al., 2018; Seifart et al., 2018; Cox et al., 2019) and the consequences of hard-to-access materials so dire that there are warnings against inadvertently creating "data graveyards, i.e. large heaps of data with little or no use to anyone" (Himmelmann, 2006, p. 4).

There are several contributing factors that result in this bottleneck with respect to transcribing minoritised languages. First, unlike major languages such as English, searchable high-quality transcriptions cannot simply be derived using a high-performance automatic speech recognition (ASR) system, whose development itself has conventionally required large quantities of transcribed speech. Second, there is typically no option for large-scale crowd-sourcing of target language transcriptions as by definition of endangerment there may be very few speakers and only a subset of speakers may be literate in this language (i.e. not the regional or national language of formal education). Third, there may not be a standardised orthography for the target language, in which case a project-specific working orthography is often developed, resulting in transcriptions with both intra- and inter-transcriber variation, requiring additional time for review and corrections. Finally, in some instances, there may also be an additional personnel bottleneck resulting from limitations on who can listen to and transcribe certain recordings (e.g. of culturally-sensitive materials such as descriptions of ceremonial procedures). Taken altogether, surveys of transcription time have reported typically requiring about 30 to 50 hours of work to transcribe one hour of recorded speech (Michaud, Castelli et al., 2014; Durantin et al., 2017; Zahrer et al., 2020).

Reviewing a century of transcription practice within linguistics, Bird (2020) argues that several parts of the transcription bottleneck are a result of a description- and analysis-centred transcription practice. In particular, he identifies three inefficient characteristics particular to conventional transcription practice: 1) transcribing phones, 2) transcribing fully, and 3) transcribing first (before translating). We review these characteristics

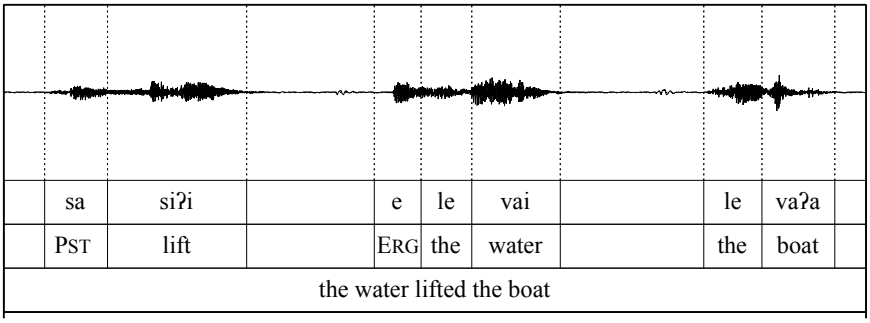

Figure 1.1: A fully transcribed and translated Samoan utterance.

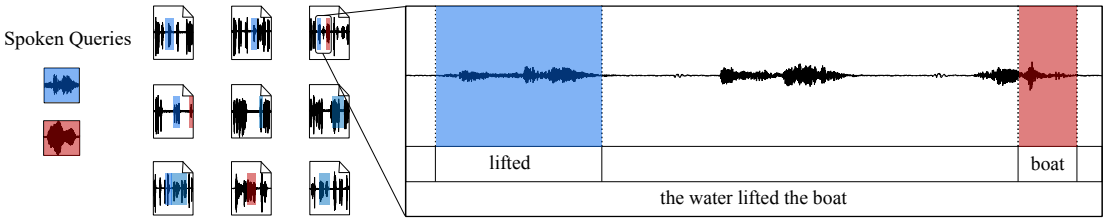

Figure 1.2: A sparsely transcribed and translated Samoan corpus and utterance.

using a Samoan utterance sourced from my Field Methods class recordings, illustrated above in in Figure 1.1. Given my goal of grammatically analysing a language largely unknown to me, I am primarily interested in textually representing the form of the spoken utterance, leading to a primary focus on the phonetic form of the whole utterance [sasiʔielevaileva?a], with the translation secondary. Using these annotations along with other data points, I can hypothesise what the various lexical and grammatical units are in Samoan and how they interact, e.g. [sa]+[siʔi] PST+lift 'lifted'.

Bird (2020) questions whether this conceptualisation of transcription effectively serves the many parties interested in working with untranscribed language documentation corpora, particularly those of unwritten, oral languages. Intended as a supplementary approach, Bird (2020) proposes 'sparse transcription' which does away with the three characteristics that contribute to the bottleneck in conventional transcription and, importantly, facilitates access to untranscribed language documentation corpora in ways that are more compatible with the skills of the speech communities. As illustrated in Figure 1.2, spoken term detection or 'word spotting' is used to facilitate searches via spoken queries (i.e. a voice-based alternative to text queries on transcriptions) and, upon confirmation of hits, speakers literate in the language of wider communication can also provide translations. In this way, the confirmed locations of various queries along with their translations allow for the corpus to be collaboratively and iteratively indexed, albeit sparsely. Additionally, this supplementary approach can help more effectively find regions of interest in the corpora to which conventional 'full' transcription efforts can be directed.

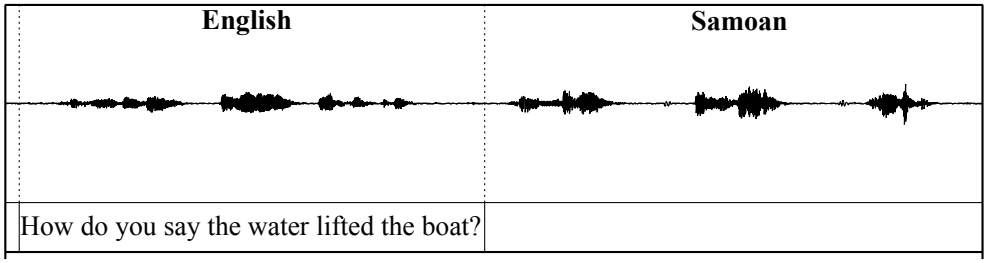

Figure 1.3: Sparse transcription via isolation and transcription of a more widely used metalanguage in a mixed-language corpus.

Following the proposal of the sparse transcription model by Bird (2020), there have been several studies investigating its potential in various workflow configurations and documentation scenarios (Le Ferrand et al., 2020; Lane & Bird, 2021; Le Ferrand et al., 2021; Lane & Bird, 2022; Le Ferrand et al., 2022). One area that has been identified for improvement is the robustness of the spoken term detection system, particularly for scenarios where the speaker of the query is different to the speakers in the corpus and when the audio of the corpus is relatively noisy. We return to this discussion below in our review of foundational models for speech and subsequent investigation of whether these models help deliver more speaker-invariant and noise-robust spoken term detection systems.

Even without a robust spoken term detection system, it may nevertheless be possible to derive a searchable index for mixed-language documentation corpora where the target language being documented is inter-mixed with a more widely spoken language for metalinguistic questions and commentary. As illustrated below in Figure 1.3, the response by the Samoan speaker was preceded by my elicitation prompt in English: *How do you say "the water lifted the boat"?* As mentioned above, it is relatively straightforward to derive searchable high-quality transcriptions using a high-performance ASR system for major languages such as English. As such, if foundation models can be used to isolate and transcribe the more widely used language, these searchable metalanguage transcriptions could provide approximate locations where certain target language words and topics are being discussed in mixed-language documentation corpora of this genre.

Ultimately sparse transcription is intended as a way to "accelerate [the work of] orthodox, contiguous transcription" (Bird, 2020, p. 737) while also facilitating immediate access to untranscribed speech corpora. If, on one hand, sparse transcription can help efficiently gather the target language transcriptions and, on the other, foundation models can be used to substantially reduce the amount of target language transcriptions required to begin ASR system development, this combination could provide a pathway for creating ASR-assisted transcription workflows for minoritised languages.

1.2 Foundation models for speech processing

Chapter 2

Conclusions

...

Appendix A

A Long Proof

...

Bibliography

Himmelmann, N. P. (1998). Documentary and descriptive linguistics.

Himmelmann, N. P. (2006). Language documentation: What is it and what is it good for. *Essentials of language documentation*, *178*(1).

Austin, P. K., & Sallabank, J. (2011). *The cambridge handbook of endangered languages*. Cambridge University Press.

Michaud, A., Castelli, E., et al. (2014). Towards the automatic processing of yongning na (sino-tibetan): Developing a'light'acoustic model of the target language and testing'heavyweight'models from five national languages. *4th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU 2014)*, 153–160.

Durantin, G., Foley, B., Evans, N., & Wiles, J. Transcription survey. In: 2017.

Foley, B., Arnold, J. T., Coto-Solano, R., Durantin, G., Ellison, T. M., van Esch, D., Heath, S., Kratochvil, F., Maxwell-Smith, Z., Nash, D., et al. (2018). Building speech recognition systems for language documentation: The coedl endangered language pipeline and inference system (elpis). *SLTU*, 205–209.

Himmelmann, N. P. (2018). Meeting the transcription challenge. In B. McDonnell, A. Berez-Kroeker & G. Holton (Eds.), *Reflections on language documentation: 20 years after Himmelmann 1998*. University of Hawai'i Press.

Seifart, F., Evans, N., Hammarström, H., & Levinson, S. C. (2018). Language documentation twenty-five years on. *Language*, *94*(4), e324–e345.

Cox, C., Boulianne, G., & Alam, J. (2019). Taking aim at the 'transcription bottleneck': Integrating speech technology into language documentation and conservation. *6th International Conference on Language Documentation and Conservation, Honolulu, HI*.

Bird, S. (2020). Sparse transcription. *Computational Linguistics*, *46*(4), 713–744. https://doi.org/10.1162/coli_a_00387

Le Ferrand, É., Bird, S., & Besacier, L. (2020). Enabling interactive transcription in an Indigenous community. *Proceedings of COLING 2020*, 3422–3428. https://doi.org/10.18653/v1/2020.coling-main.303

Zahrer, A., Žgank, A., & Schuppler, B. (2020). Towards building an automatic transcription system for language documentation: Experiences from muyu. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2893–2900.

Lane, W., & Bird, S. (2021). Local word discovery for interactive transcription. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2058–2067.

Le Ferrand, É., Bird, S., & Besacier, L. (2021). Phone based keyword spotting for transcribing very low resource languages. *Proceedingd of the 19th Workshop of the Australasian Language Technology Association (ALTA) 2021*, 79–86.

Holton, G., Leonard, W. Y., & Pulsifer, P. L. (2022, January). Indigenous Peoples, Ethics, and Linguistic Data. In *The Open Handbook of Linguistic Data Management*. The MIT Press. https://doi.org/10.7551/mitpress/12200.003.0008

Lane, W., & Bird, S. (2022). A finite state aproach to interactive transcription. *Proceedings of the first workshop on NLP applications to field linguistics*, 1–10.

Le Ferrand, É., Bird, S., & Besacier, L. (2022). Learning from failure: Data capture in an australian aboriginal community. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4988–4998.

Berez-Kroeker, A. L., Gabber, S., & Slayton, A. (2023). Recent advances in technologies for resource creation and mobilization in language documentation. *Annual Review of Linguistics*, *9*, 195–214.